

TECHNICAL INFORMATION

*“Ensuring Appropriate Assessment
for All Students with Special Needs”*

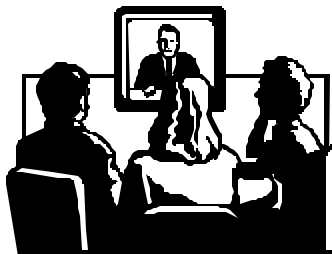
Wednesday, September 29, 2004

TIME: 1:00 p.m. – 3:00 p.m. EDT
12:00 p.m. – 2:00 p.m. CDT
11:00 a.m. – 1:00 p.m. MDT
10:00 a.m. – 12:00 p.m. PDT

TEST TIME: 12:30 p.m. – 1:00 p.m. EDT
11:30 a.m. – 12:00 p.m. CDT
10:30 a.m. – 11:00 a.m. MDT
9:30 a.m. – 10:00 a.m. PDT


SATELLITE: IA-6 (formerly TELSTAR – 6)
BAND: C-BAND
TRANSPONDER: 4
CHANNEL: 4
POLARITY: HORIZONTAL
AUDIO: 6.2 / 6.8 MHz
LOCATION: 93° WEST LONGITUDE
FREQUENCY: 3780 MHz

TECHNICAL TROUBLE NUMBER (Day of the program only)
Pittsburgh International Teleport (TV Operations Center) - 800-634-6530




Slide 1

Alternate assessment, gaps, and other challenges!




A view of current practices from the technical assistance perspective

Panelist: Rachel Quenemoen, technical assistance team leader, NCEO




National Center on Educational Outcomes Fall, 2004




Slide 2

Overview

- A brief review of regulation language on alternate assessment
- A close look at options for alternates, and how the term “grade level” plays out in state discussions
- Clarification of academic content and achievement standards – definitions and relationships
- Examples of students with significant cognitive disabilities achieving in the grade level curriculum: Massachusetts; Kentucky
- Assumptions about “other” students who may be affected by gaps in instruction, curriculum, and assessment
- State illustrations: Connecticut; Ohio




National Center on Educational Outcomes Fall, 2004




Slide 3

Alternate Assessments as defined in “1% Rule”

- Aligned with the State’s content standards.
- Yield results separately in reading/language arts and math.
- Designed and implemented to support use of the results to determine AYP.





National Center on Educational Outcomes Fall, 2004



Slide 4

Alternate Assessments should have...

- Clearly defined structure
- Guidelines for which students may participate
- Clearly defined scoring criteria and procedures
- Report format that clearly communicates student performance in terms of the academic achievement standards defined by the State

National Center on Educational Outcomes Fall, 2004



Slide 5

Alternate Assessments

Must meet the same requirements for high technical quality that apply to regular assessments under NCLB:

- Validity
- Reliability
- Accessibility
- Objectivity
- Consistent with nationally-recognized professional and technical standards.



See Peer Review Guidance, AERA papers, other technical resources

National Center on Educational Outcomes Fall, 2004

Slide 6

States may use more than one alternate assessment

- Alternate assessment scored against grade-level achievement standards
- Alternate assessment scored against alternate achievement standards
- Both must be aligned to the State's academic content standards



National Center on Educational Outcomes Fall, 2004

Slide 7

Alternate What?

Content standards are not changed for any of the options – now that we have grade level definitions, we start there for all assessment options.

Alternate achievement standards may be set for students with significant cognitive disabilities.



National Center on Educational Outcomes Fall, 2004

Slide 8

What does “grade level” mean?

See handout on grade level

Starting Point: Access to and progress in the SAME challenging curriculum/ content for ALL students, including those with significant disabilities

National Center on Educational Outcomes Fall, 2004

Slide 9

Title I Regulations Addressing Academic Content and Achievement Standards

- July 5, 2002
- December 9, 2003

National Center on Educational Outcomes Fall, 2004


Slide 10

Grade level terms, continued:


Grade level achievement standards

Alternate achievement standards

Alternate achievement standards set by grade level (“grade level alternate achievement standards” or “grade by grade alternate achievement standards”)




National Center on Educational Outcomes Fall, 2004




Slide 11

Assessment Options

- General assessment w/wo accommodations
- Alternate assessment on grade level achievement standards –**alternate ways of showing proficiency on the grade-level content standards (or GLEs) against grade-level achievement standards**
- Alternate assessment on alternate achievement standards – **alternate ways of showing proficiency on the SAME grade-level content standards (or GLEs) (extended or expanded) against alternate achievement standards**



National Center on Educational Outcomes Fall, 2004




Slide 12


Development Process for Alternate Achievement Standards

- Begin from state academic content standards for grade in which student is enrolled
- Adapt or extend standards to ensure meaningful access for students with significant cognitive disabilities
- Draft *proposed* achievement descriptors (sometimes called “Performance Level Descriptors”) that describe at least three proficiency levels
- Assess students and score results against preliminary scoring criteria and draft achievement descriptors
- Set achievement standards, refine achievement descriptors, fine-tune the assessment method and scoring criteria

Adapted from Jan Sheinker presentation on April 2004 teleconference - <http://education.mn.edu/news/Presentations/telech.htm>







National Center on Educational Outcomes Fall, 2004



Slide 13

ACHIEVEMENT STANDARDS

Performance levels		Labels each level of achievement
Performance descriptors		Describes each level of performance
Exemplars		Sample student work at each level of performance
Cut Scores		Scores that separate the different levels of performance



FROM: Handbook for Professional Development in Assessment Literacy, Jan Shinkre and David Ruffield

[illegible]

Slide 14

What does it look like when students with significant cognitive disabilities access and make progress in the general curriculum at grade level?

- Massachusetts – Bobby J and the life cycle of the frog
– Dan Wiener, Massachusetts Department of Education
- Kentucky – video clips – Jacqui Kearns, University of Kentucky ILSSA Inclusion Project

 National Center on Educational Outcomes Fall, 2004 

Slide 15


Who are the “other” students who may be affected by a gap of some kind?

Common question: How many students “can” achieve grade-level achievement standards, with the best instruction and access?

Kevin McGrew studies and NCEO paper:
<http://www.japsych.com/index.htm>


National Center on Educational Outcomes Fall, 2004

Slide 16




Alternative question:

How many schools **currently ensure every child has the services, supports, and specialized instruction necessary to succeed in the grade-level curriculum?**




National Center on Educational Outcomes Fall, 2004




Slide 17

Assumption: A few students in the “gap” truly cannot show what they know on our current assessments. But others – perhaps many - students have been failed by our system of curriculum and instruction. They have not been taught the challenging grade-level curriculum. The policy goal is first and foremost to correct that situation.

Assumption: Some (unknown number) students in the “gap” may not achieve to proficiency at grade level by high school, even with the best possible curriculum and instruction, but we don’t know which ones or how many. We need to find that out by giving them the opportunity to succeed.



National Center on Educational Outcomes Fall, 2004




Slide 18


Solutions?

Assumption: We need to close gaps in curriculum and instruction as well to make our assessment system truly accessible – and really push practice to make that occur as quickly as possible!

Assumption: All students, including those who we may not have expected to achieve in the past, have the right to be taught as if they can succeed, even if they all do not ultimately achieve proficiency in all areas.





National Center on Educational Outcomes Fall, 2004



Slide 19

What can we do in our assessment and accountability policies and practices to move MOST (99%) students into the general assessment, and to ensure all students achieve at the highest level possible?


Connecticut will address that question in their presentation!




National Center on Educational Outcomes Fall, 2004

Slide 20

What would a schooling system built on grade-level content for ALL students look like?



How do we ensure this is happening in OUR schools? Ohio will address that question in their presentation!




National Center on Educational Outcomes Fall, 2004

Slide 21

Strand: Life Science (Biology)
Learning Standards for ... Life Cycles and Heredity
Grade 3 (MA)

Recognize that plants and animals go through predictable life cycles that include birth, growth, development, reproduction, and death.

Describe the major stages that characterize the life cycle of the frog and the butterfly as they go through metamorphosis




National Center on Educational Outcomes Fall, 2004


Slide 22

**Essence of the Standards – Bobby J
Student Work from MA**

- Recognize the 4 major stages of an organism's life cycle:
 - Birth
 - Development/growth
 - Reproduction
 - Death
- Explain frog/butterfly life cycles




National Center on Educational Outcomes Fall, 2004




Slide 23

**Teaching and learning in the grade-
level content - KY**

Kentucky video clips – Inclusion Project



National Center on Educational Outcomes Fall, 2004



Different ways the words “grade level” play out in the discussion of alternate achievement standards – some ideas for discussion based on current practices.

Access to general curriculum based on state content standards/expectations for ALL students, including those with significant cognitive disabilities:

ALL students must have access to and make progress in the same challenging curriculum based on state grade level content standards (or grade level expectations if defined that way), including students with significant cognitive disabilities. The approach of adapting and accommodating student learning to allow development of knowledge and skills in the grade level curriculum is proving to unlock learning for students who have never had access to the challenging content before, dramatically increasing expectations and outcomes. States will need to have thoughtful discussion about how to ensure that students with significant cognitive disabilities have access to the essence of content standards (or GLEs) as extended from the grade level, and not simply generic low level skills that “extend” universally regardless of grade level.

Grade level achievement standards:

Definitions of achievement (or performance) at varying levels (typically 3 or more) that define grade level achievement on the grade level academic content as proficient or not – set on general assessment, must be the same for alternate assessment on grade level achievement standards.

Alternate achievement standards:

Definitions of achievement (or performance) at varying levels (typically 3 or more) that define how students with significant cognitive disabilities demonstrate proficiency on the grade level academic content. These alternate definitions of achievement typically are qualitatively and quantitatively different from grade level achievement standards – and states must conduct and be able to defend a very thoughtful process for defining these to reflect appropriate, challenging, and high expectations for achievement in the content domains as defined for the student’s enrolled grade. In some states, this includes how students use accommodations, modifications, and assistive technology in order to learn and then demonstrate skills and knowledge aligned to the grade level content, or descriptions of how students are demonstrating access and entry skills in the context of grade level content and curricular activities.

Alternate achievement standards set by grade level (occasionally referred to as “grade level alternate achievement standards” or “grade by grade alternate achievement standards”)

Alternate achievement standards should demonstrate increasing complexity, progress, or appropriate levels of challenge for each increasing grade level. That MAY mean setting of separate alternate achievement standards at each grade level or it MAY mean clarifying how increasing grade level cut scores mean increasing expectations over the grade levels as part of the achievement level descriptors. Each state needs to grapple with this issue in order to ensure high expectations at each grade level and to show progress as students move up the grade levels. This is a complex concept – and will need extended discussion and policy review in each state to ensure that the policy goals of appropriate challenge and high expectations are achieved over the grades.

All of this requires thoughtful discussion in a state of how students with significant cognitive disabilities access grade level content, and how they build and demonstrate proficiency on that content. If the state ensures alignment of alternate assessments to enrolled grade level content for every student, grade level content which by definition increases in complexity grade by grade, it will be easier to defend how the alternate achievement standards set on these alternate assessments also increase in complexity grade by grade.

See attached citations from Title I Regulations defining academic content and achievement standards, July 5, 2002, and December 9, 2003.

TWO CITATIONS FROM TITLE I REGULATIONS DEFINING ACADEMIC CONTENT AND ACHIEVEMENT STANDARDS: July 5, 2002 and December 9, 2003

[Federal Register: July 5, 2002 (Volume 67, Number 129)]
[Rules and Regulations]
[Page 45037-45047]
From the Federal Register Online via GPO Access [wais.access.gpo.gov]
[DOCID:fr05jy02-12]

Sec. 200.1 State responsibilities for developing challenging academic standards.

(a) Academic standards in general. A State must develop challenging academic content and student academic achievement standards that will be used by the State, its local educational agencies (LEAs), and its schools to carry out subpart A of this part. These academic standards must--

(1) Be the same academic standards that the State applies to all public schools and public school students in the State, including the public schools and public school students served under subpart A of this part;

(2) Include the same knowledge, skills, and levels of achievement expected of all students; and

(3) Include at least mathematics, reading/language arts, and, beginning in the 2005-2006 school year, science, and may include other subjects determined by the State.

(b) Academic content standards. (1) The challenging academic content standards required under paragraph (a) of this section must--

(i) Specify what all students are expected to know and be able to do;

(ii) Contain coherent and rigorous content; and

(iii) Encourage the teaching of advanced skills.

(2) A State's academic content standards may--

(i) Be grade specific; or,

(ii) Cover more than one grade if grade-level content expectations are provided for each of grades 3 through 8.

(3) At the high school level, the academic content standards must define the knowledge and skills that all high school students are expected to know and be able to do in at least reading/language arts, mathematics, and, beginning in the 2005-06 school year, science, irrespective of course titles or years completed.

(c) Academic achievement standards. (1) The challenging student academic achievement standards required under paragraph (a) of this section must--

(i) Be aligned with the State's academic content standards; and

(ii) Include the following components for each content area:

(A) Achievement levels that describe at least--

(1) Two levels of high achievement--proficient and advanced--that determine how well students are mastering the material in the State's academic content standards; and

(2) A third level of achievement--basic--to provide complete information about the progress of lower-achieving students toward mastering the proficient and advanced levels of achievement.

[[Page 45040]]

(B) Descriptions of the competencies associated with each achievement level.

(C) Assessment scores ("cut scores") that differentiate among the achievement levels as specified in paragraph (c)(1)(ii)(A) of this section, and a description of the rationale and procedures used to determine each achievement level.

(2) A State must develop academic achievement standards for every grade and subject assessed, even if the State's academic content standards cover more than one grade.

(3) With respect to academic achievement standards in science, a State must develop--

(i) Achievement levels and descriptions no later than the 2005-06 school year; and

(ii) Assessment scores ("cut scores") after the State has developed its science assessments but no later than the 2007-08 school year.

(d) Subjects without standards. If an LEA serves students under subpart A of this part in subjects for which a State has not developed academic standards, the State must describe in its State plan a strategy for ensuring that those students are taught the same knowledge and skills and held to the same expectations in those subjects as are all other students.

(e) Other subjects with standards. If a State has developed standards in other subjects for all students, the State must apply those standards to students participating under subpart A of this part.

(Authority 20 U.S.C. 6311(b)(1))

(Approved by the Office of Management and Budget under control number 1810-0576)

[Federal Register: December 9, 2003 (Volume 68, Number 236)]
[Rules and Regulations]
[Page 68697-68708]
From the Federal Register Online via GPO Access [wais.access.gpo.gov]
[DOCID:fr09de03-27]

[[Page 68697]]

The Secretary amends part 200 of title 34 of the Code of Federal Regulations as follows:

PART 200--TITLE I--IMPROVING THE ACADEMIC ACHIEVEMENT OF THE DISADVANTAGED

0

1. The authority citation for part 200 continues to read as follows:

Authority: 20 U.S.C. 6301 through 6578, unless otherwise noted.

0

2. In Sec. 200.1, revise paragraph (a)(1), redesignate paragraphs (d) and (e) as (e) and (f), and add new paragraph (d) to read as follows:

Sec. 200.1 State responsibilities for developing challenging academic standards.

(a) * * *

(1) Be the same academic standards that the State applies to all public schools and public school students in the State, including the public schools and public school students served under subpart A of this part, except as provided in paragraph (d) of this section;

* * * * *

(d) Alternate academic achievement standards. For students under section 602(3) of the Individuals with Disabilities Education Act with the most significant cognitive disabilities who take an alternate assessment, a State may, through a documented and validated standards-setting process, define alternate academic achievement standards, provided those standards--

(1) Are aligned with the State's academic content standards;

(2) Promote access to the general curriculum; and

(3) Reflect professional judgment of the highest achievement standards possible.

* * * * *

0

3. In Sec. 200.6, revise paragraph (a)(2)(ii) and add new paragraph (a)(2)(iii) to read as follows:

Sec. 200.6 Inclusion of all students.

* * * * *

(a) * * *

(2) * * *

(ii)(A) Alternate assessments must yield results for the grade in which the student is enrolled in at least reading/language arts, mathematics, and, beginning in the 2007-2008 school year, science, except as provided in the following paragraph.

(B) For students with the most significant cognitive disabilities, alternate assessments may yield results that measure the achievement of those students relative to the alternate academic achievement standards the State has defined under Sec. 200.1(d).

(iii) If a State permits the use of alternate assessments that yield results based on alternate academic achievement standards, the State must--

(A)(1) Establish and ensure implementation of clear and appropriate guidelines for Individualized Educational Program (IEP) teams to apply in determining when a child's significant cognitive disability justifies assessment based on alternate academic achievement standards; and

(2) Ensure that parents of those students are informed that their child's achievement will be based on alternate achievement standards; and

(B) Report separately, under section 1111(h)(4) of the ESEA, the number and percentage of students with disabilities taking--

(1) Alternate assessments based on the alternate academic achievement standards defined under Sec. 200.1(d);

(2) Alternate assessments based on the academic achievement standards defined under Sec. 200.1(c); and

[[Page 68703]]

(3) Regular assessments, including those administered with appropriate accommodations.

(C) Document that students with the most significant cognitive disabilities are, to the extent possible, included in the general curriculum and in assessments aligned with that curriculum;

(D) Develop, disseminate information on, and promote use of appropriate accommodations to increase the number of students with the most significant cognitive disabilities who are tested against grade-level academic achievement standards; and

(E) Ensure that regular and special education teachers and other appropriate staff know how to administer assessments, including making appropriate use of accommodations, for students with the most significant cognitive disabilities.

* * * * *

0

4. In Sec. 200.13, revise the introductory text of paragraph (b) and paragraph (b)(1), redesignate paragraph (c) as paragraph (d), and add new paragraph (c) to read as follows:

PROPOSED AMENDMENTS September 2004 (Middle column)

**I Say Potato, You Say Potahto:
The Assessment-speak Gap
Between General and Alternate Assessment Experts**

A SIDE-BY-SIDE GLOSSARY

**Joseph M. Ryan
Arizona State University West**

**Rachel F. Quenemoen and Martha L. Thurlow
National Center on Educational Outcomes
University of Minnesota**

April 15, 2004

Paper presented at the annual meeting of the American Educational Research Association (AERA). It was prepared, in part, by the National Center on Educational Outcomes through a Cooperative Agreement (#H326G000001) with the Research to Practice Division, Office of Special Education Programs, U.S. Department of Education. Points of view or opinions expressed in the paper are not necessarily those of the U.S. Department of Education, or Offices within it.

Assessment Term, Concept, or Procedure	Traditional connotations as used in assessing students with significant cognitive disabilities	<i>Meeting in the middle: Building on the expertise of all partners</i> <i>Implications, insights, and inspiration</i>	Traditional connotations as used in assessing students in the general education population
Population	<p>Very small group of students (dozens in a state)</p> <p>State-to-state variation of students who take alternate assessment/aas, multiple alternate assessments/aas, pressure from 1% Rule</p> <p>This is a highly variable population in terms of learner characteristics, available response repertoires, and often competing complex medical conditions</p> <p>“Outliers” can be a large proportion of this very small population</p>		<p>Tens or hundreds of thousands of students</p> <p>Rules for inclusion and exclusions vary across time and setting</p> <p>Often homogeneous in the aggregate with respect to what is being measured (e.g., the construct has the same meaning for most students although students may vary in amount of knowledge/skill).</p> <p>“Outliers” who are not homogeneous are a relatively small proportion of the large population</p>

Construct domain	<p>The applicable construct domains for students are often defined individually, through flexible access points to the grade-level content</p> <p>Generally refers to observable behaviors related to performance of content related skills and knowledge</p> <p>May be defined through IEP process in states that are training on standards-based IEPs; may involve progress on standards-based IEP goals</p> <p>No consensus theory of learning as yet exists in the academic content areas for these children, that is, what patterns of growth they show on the path to competence</p>		<p>State standards generally define grade-level construct domains all students</p> <p>Defines learning targets in terms of content, cognitive processes, and performance</p> <p>Usually emphasizes content, also refers to cognition, e.g., remembering, comprehending, applying, and more complex processes</p> <p>Specifies the boundaries (what's in and what's not), structures, and relationships among elements</p>
Assessment Format: Tests and Items	<p>The majority of states use portfolio, body-of-evidence, or other performance-based models for their alternate assessments on alternate achievement standards (aas)</p>		<p>Test are generally given under standard conditions in terms of content, format, timing, and response mode</p> <p>A common test blueprint is</p>

	<p>Most alternate assessments/aas include relatively few open-ended tasks that are often tailored to the individual student.</p> <p>In many states, teachers are trained to design assessment tasks to demonstrate student knowledge and skills, often embedded in ongoing instruction</p>		<p>used across test forms</p> <p>Item formats vary widely and include selected response, short answer, extended open-ended task and/or response, complex constructed response, and performance assessments</p> <p>Item formats vary to reflect the different learning objective being assessed</p>
Generalization / Generalizability	<p>Given the limited understanding of the construct domain, and lack of consensus on a theory of learning in the academic content for these students, and the varying coverage of the domain, generalization as traditionally defined in measurement is a challenge</p> <p>The term “generalization” is a foundational term used by special educators, and is a common scoring criterion, meaning: <i>Student performance of skills or knowledge learned in one</i></p>		<p>Assessments should provide representative coverage of the construct domain content and processes so that score interpretation is not limited to the sampled tasks on the specific assessment</p> <p>Generalizability is usually considered an aspect of validity although the “consistency” connotation reflects the concept of reliability</p> <p>Generalizability studies are rarely part of local assessment programs and are not always</p>

	<i>setting or for one purpose is evidenced in additional settings or for different purpose</i>		included in state programs.
Reliability	<p>Often refers to whether a student can demonstrate the same behavior two or three times, or through triangulated data sources</p> <p>Cannot be easily quantified in terms of classical test theory concepts of true and error scores</p> <p>Some states report inter-rater reliability statistics as one indicator of reliability for alternate assessments. Although reporting the consistency of scoring processes is valuable, reporting inter-rater agreement statistics as if they are reliability coefficients is misleading</p>		<p>Usually refers to consistency in response to items, which are viewed as sampled replications from a construct domain</p> <p>Used to evaluate inferences about the likelihood that students would perform similarly on the same or parallel form of the assessment</p> <p>Easily quantified in indices of internal consistency, alternative form, tests-retest reliability</p>
Error of Measurement	Very difficult to index because of small sample sizes and narrowly defined behavioral domains		Provides a quantification of the amount of error that can be expected in students' scores

			<p>Used to establish confidence intervals or bands within which students' true scores are known with a specified level of probability</p> <p>Straight forward in both classical and IRT approaches</p>
Validity	<p>Some validity studies have looked at the process used in alternate assessment design in states, specifically around defining the scoring criteria. Stakeholder agreement on criteria reflecting achievement for students with significant disabilities then shapes the design of the alternate assessment.</p> <p>A few studies have looked at concurrent validity of alternate assessment scores against other measures of quality programming and outcomes for students with significant disabilities</p> <p>Current work is being done</p>		<p>An integrated evaluative judgment about the degree to which evidence and theory support the adequacy and appropriateness of inferences and actions based on assessment information (Messick, 1989, p 13)</p> <p>In most settings, validity rest largely on demonstrating that the assessment reflects the content standards it is designed to measure.</p> <p>The degree to which items reflect the content standards is usually assessed by a content review panel.</p> <p>Evidence about adequate care</p>

	<p>on content validity or at a minimum, alignment of extended standards to general standards, and ultimately to the alternate assessment.</p> <p>Documentation of adequate care and implementation of recognized procedures in setting of alternate achievement standards has occurred in a few states. There is limited understanding in special education of what setting standards involves, what it means</p> <p>Correlational studies have documented rapid shifts in instruction and curriculum in the desired directions in several states through teacher surveys and observational protocols</p>		<p>and implementation of recognized procedures in the item and test development processes often is used as validity evidence</p> <p>Evidence about adequate care and implementation of recognized procedures in the setting performance standards often is used as validity evidence</p> <p>Correlations with external variables (convergent and divergent) are frequently used as validity evidence</p>
Fairness	A layman's version of a fairness discussion is a common aftermath to the		Often seen as an aspect of validity

	<p>first year of alternate assessment/aas. These discussions focus on whether these assessments measure the skill of the teacher or the skill of the student, whether scoring processes are of high quality and are applied consistently, and whether it is appropriate or desirable to expect these students to learn academic content</p> <p>Generally, the discussion is focused on how unfair the new assessments are to teachers. States respond with additional training support in many cases, although some states have reduced requirements considerably in the face of the outcry</p> <p>The accountability requirements of NCLB may change the nature of fairness discussions</p> <p>Proponents of alternate assessments/aas suggest that OTL is the major fairness</p>		<p>Deals specifically with evaluating assessments for bias, meaning that tests scores are influenced by factors irrelevant to the construct being measured</p> <p>Generally examined through studies of differential item functioning (DIF)</p> <p>Bias-sensitivity panels review assessment items and task for any offensive features and for opportunity to learn (OTL) as a standard element in test development</p> <p>Sources of construct irrelevant variance (e.g., language skills in math or social studies) are also examined judgmentally and empirically</p> <p>OTL is an aspect of fairness that is examined in some assessment programs</p> <p>Assessment data are disaggregated and the validity of the assessment for each</p>
--	--	--	--

	issue for this group of children		subgroup is considered
Test Administration	<p>These assessments tend to be individually tailored to the response repertoire of the individual student. The content, items, format, timing, and response mode are all individualized</p> <p>Level of challenge is a criterion in several states on which the evidence is scored</p> <p>A few states have developed common tasks with flexible modes of response, scoring on level of prompting needed before a student can respond</p>		<p>The critical feature of test administration is that tests are generally given under standard conditions in terms of content, items, format, timing, and response mode</p> <p>In most cases, students take exactly the same test or a form that is equivalent in content and difficulty</p> <p>In a few instances, like NAEP, students take a subset or sample of items but in such cases individual scores are not reported</p> <p>Amount of time student have to take the tests may vary from a fixed period to un-timed conditions</p>
Scoring	Performance assessments are scored against carefully developed standards-referenced rubrics applied by trained raters in many states.		<p>Selected-response questions are machine-scored against a key</p> <p>Short answer, extended</p>

	<p>The scoring rubrics reflect the task and content domain structure and, thus, are part of the validity evidence</p> <p>The raters are trained to a mastery criterion and then check papers, read behinds, and rater agreement indices are employed to monitor scoring</p> <p>Some states have regional certified scorers administer the tasks or checklist, or they may document the evidence supporting teacher scoring in a sample of cases</p> <p>Other states permit teacher scoring and reporting of student performance. Some require a sample audit; others rely on teacher judgment</p>		<p>response, and other performance assessments are scored against carefully developed standards-referenced rubrics applied by trained raters</p> <p>The scoring rubrics reflect the task and content domain structure and, thus, are part of the validity evidence.</p> <p>The raters are trained to a mastery criterion and then check papers, read behinds, and rater agreement indices are employed to monitor scoring</p>
Interpretation	<p>In a few states, student performance is interpreted relative to achievement standards resulting in students being classified into various achievement levels</p>		<p>Student performance is interpreted normatively (percentiles, stanines, etc)</p> <p>Student performance is interpreted relative to</p>

	<p>Achievement standards are generally based on panel review of score patterns and student work, and cutscores are selected using various recognized procedures</p> <p>Achievement levels often have substantively rich descriptions that aid in interpretation</p> <p>Concern focuses on “How high is high enough,” challenge, appropriateness</p>		<p>performance standards resulting in students being classified into various achievement levels</p> <p>Performance standards are generally based on an examination of item content when cutscores are selected using various recognized procedures are employed</p> <p>Achievement levels often have substantively rich descriptions that aid in interpretation</p>
Consequence	<p>Consequential validity is the primary area of study of the effects of alternate assessment on alternate achievement standards.</p> <p>Correlational studies have documented rapid shifts in instruction and curriculum in the desired directions in several states through teacher surveys and observational protocols.</p> <p>These are students who in</p>		<p>Often incorporated as an aspect of validity</p> <p>Involves examining the intended and unintended consequences of the intended assessments use</p> <p>Not always evaluated</p> <p>The impact of an assessment applications that has a specific purpose (e.g., identify students in need of remediation) should be</p>

	many cases have had no access to the general curriculum.		<p>examined to see if the impact is achieved (e.g., did the students receive remediation)</p> <p>Assessments designed to yield information to be used in educational decisions should be examined to determine what, if any, role the results play in decisions making.</p> <p>Unintended outcomes should be examined to determine if they are related to characteristics of the students that are not related to the construct being measured.</p>
--	--	--	---

I Say Potato, You Say Potahto: An AERA Conference Discussion Paper

**Rachel F. Quenemoen and Martha L. Thurlow
National Center on Educational Outcomes
University of Minnesota**

April 15, 2004

The authors acknowledge the substantial contributions made to this paper by Scott Marion of the National Center for Improvement of Educational Assessment (NCIEA) in the section on how traditional approaches to technical adequacy apply; and the substantial contributions of Jacqui Kearns of ILSSA at the University of Kentucky in the overall conceptual framework of the paper.

Paper presented at the annual meeting of the American Educational Research Association (AERA). It was prepared, in part, by the National Center on Educational Outcomes through a Cooperative Agreement (#H326G000001) with the Research to Practice Division, Office of Special Education Programs, U.S. Department of Education. Points of view or opinions expressed in the paper are not necessarily those of the U.S. Department of Education, or Offices within it.

I Say Potato and You Say Potahto: An AERA Conference Discussion Paper

Alternate assessments developed to assess students with significant cognitive disabilities are relatively new in most states, developed for students who were not included in most large-scale assessments until Federal law mandated their participation. The requirement for states to develop these assessments first appeared in the Individuals with Disabilities Education Act Amendments of 1997 (IDEA 97). The No Child Left Behind Act of 2001 (NCLB) included the results of these assessments in its accountability requirements, and NCLB regulations clarified that students participating in alternate assessments could be held to alternate achievement standards (December 2003 Title I Regulations).

To meet Federal accountability purposes, states and testing companies have struggled to identify technically adequate and educationally sound methods of assessing this small group of students with significant cognitive disabilities. Typically, both experts in educational programming for these students and key stakeholders have advised state assessment offices in defining what the best possible outcomes of standards-based instruction should be for the students. From those definitions, states and test company partners have developed assessments to measure the outcomes for school, district, and state accountability purposes. Most states have then worked with their technical advisory committees (TAC) to discuss whether the methods meet basic standards of technical adequacy, often through review and comment on the state's technical manual for the assessment.

In the next year, state assessment systems will undergo Title I peer review to determine whether the systems meet the requirements of NCLB. Technical manuals and TAC input will be important pieces of documentation. Yet not many states or testing companies are as confident that they understand what is necessary to document these alternate assessments as they are for the general assessment. Technical experts have raised concerns that many of these approaches do not "fit" traditional models, and seem to have questionable alignment to the Joint Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999). As one TAC member said after review of a proposed body of evidence approach, "This may be a fine way of looking at classroom student work for this group of students, but it isn't measurement for accountability purposes." TAC members, and even the test company psychometricians responsible for producing technical manuals to present to the TACs, are uncomfortable with the limited tools available to understand what is occurring. And, in turn, experts in educational programming for these students and key stakeholders who have advised states on alternate assessments are baffled at what they perceive to be reluctance by measurement experts to "take these assessments (and by inference, these children) seriously."

Purpose of this paper. In the past five years, special education and educational measurement experts have attempted to learn one another's "culture and language" as we have partnered to build assessments that measure the achievement of every student. We have struggled in our efforts. Alternate assessments of the students with significant cognitive disabilities have posed particular challenges to these partnerships.

This paper is a companion piece to a side-by-side annotated glossary of terms in measurement language for students with significant cognitive disabilities and in measurement language for students in the general assessment population. The glossary was developed through a cross-disciplinary partnership, given evidence that at times we, that is, special education and measurement experts, are using the same terms with very different connotations (Ryan, Quenemoen, & Thurlow, 2004). The glossary includes the terms population, construct domain, assessment format (tests and items), generalization/generalizability, reliability, error of measurement, validity, fairness, test administration, scoring, interpretation, and consequence and is available from the authors of this paper.

In this paper, we discuss the current status of development of alternate assessments for students with significant cognitive disabilities, how traditional approaches to technical adequacy apply to these assessments, and why we should commit to cross-disciplinary work to improve these assessments, including partnering to reconceptualize traditional measurement terms if necessary. Finally, we propose how we can work together to advance our mutual ability to build assessments that work for all students.

Current status. Since alternate assessments were first required to be operational in 2000, researchers have documented state approaches, most typically portfolio or body of evidence methods, but also including performance assessments and checklists (Thompson & Thurlow, 2001). Regardless of the approach used for alternate assessment, several steps have been identified where both technical adequacy and educational soundness must be carefully addressed. The methods used in states to extend or expand the state content standards for the purpose of aligning alternate assessments to the same academic content as the general assessments are an essential step, studied by many researchers (Browder, 2001; Browder, Flowers, Ahlgrim-Delzell, Karvonen, Spooner, & Algozzine, 2002; Kleinert & Kearns, 2001; Thompson, Quenemoen, Thurlow, & Ysseldyke, 2001; Tindal, in press). Although the academic content covered must be aligned to the same content standards as the general assessment, researchers are identifying multiple ways states are defining the constructs being measured, based on professional understanding of how this very small group of the most challenged students demonstrates successful academic learning.

Additional thoughtful development is necessary to clarify how learning in the content is shown by these students. There is not as yet consensus on a theory of learning in the academic content for these students, although states often address what state stakeholders believe about their learning through the criteria used to score alternate assessment responses or evidence (Quenemoen, Thompson, & Thurlow, 2003). These efforts build on literature defining successful outcomes for students with significant cognitive disabilities (Kleinert & Kearns, 1999; Ysseldyke & Olsen, 1997). Yet, because of the new demands of federal and state laws requiring increased technical adequacy, these efforts must result in precise definitions of what we are measuring as we look at achievement for students with the most significant (typically cognitive or multiple) disabilities (Quenemoen et al., 2003). Other researchers have begun defining how to document the validity and reliability of these assessments (Garrett, Towles, Kleinert, & Kearns 2003; Kearns & Kleinert, 1999; Turner, Baldwin, Kleinert, & Kearns, 2000; White, Garret, Kearns, Grisham-Brown, 2004), although most states have not as yet done so. Finally, there is emerging literature on standard-setting approaches that can be used for

alternate assessment in order to define what "proficient" means for accountability purposes (Arnold, 2003; Olson, Mead, & Payne, 2002; Roeber, 2002; Weiner, 2002).

Typically, a state has its assessment system TAC members, often measurement experts from universities and national centers, review proposed assessments for technical adequacy. Very few TAC members have had any previous experience or contact with the achievement of students with significant cognitive disabilities, and have struggled to understand the proposals that come to them. By contrast, many of the special education expert advisors who serve as experts to states on the academic performance of the students with significant cognitive disabilities have had limited experience with measurement for large-scale assessment and accountability purposes, and have struggled to understand the technical concerns of the TAC members. This communications challenge has limited the ability of either group—measurement expert and special education expert—to articulate key concerns and collaboratively resolve them in ways that benefit the students. Curriculum experts are overlooked in the discussion, resulting in confusion about just what is being measured. Yet, the alliance of all three partners is necessary to ensure a technically adequate and educationally sound assessment that can result in improved outcomes for these students.

How do traditional approaches to technical adequacy apply? Many writers of technical reports for general assessments attempt to align their analyses and results with the Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999), particularly when there are student or school stakes requiring that the inferences drawn from the assessment be valid, reliable and fair (AERA, APA, & NCME, 1999). This is an obvious and important first step, but often not fully met. Leading measurement theorists (e.g., Cronbach, Messick), including the authors of the 1985 and 1999 standards for educational measurement, are clear that validity is the most important technical criterion for educational assessment. In order to address validity, test developers must have a clear understanding of both the target constructs and how students with significant cognitive disabilities are expected to come to know these constructs, a clear understanding of the theory of learning for these students in the academic domains (Pellegrino, Chudowsky, & Glaser 2001).

The majority of states use portfolio or other performance-based models for their alternate assessments. Over a decade ago, as performance-based assessments started to become more widely used with students in the general population, several theorists started to question and offer solutions for evaluating the technical adequacy of these "new" assessment types (e.g., Linn, Baker, & Dunbar, 1991; Messick, 1995; Moss, 1992). While not parallel, there are several analogous challenges in that validity and reliability needed some degree of reconceptualization in order to be useful for evaluating the quality of performance-based assessments. We are not suggesting that this reconceptualization has been entirely successful or complete but we believe that addressing issues in alternate assessment can help shed light on these types of concerns related to assessments for students in the general population.

Reliability is often mentioned in the same breath as validity as the other essential technical quality. In fact, a common saying in educational measurement is that "you cannot have validity without reliability." This is certainly true from a traditional perspective, but perhaps it will be necessary to move beyond these traditional perspectives in the context of

alternate assessment. In a recent special issue of *Educational Measurement: Issues and Practices*, the authors offered approaches for reconceptualizing traditional measurement criteria so they can be useful for evaluating the quality of classroom assessment system. In particular, Jeffrey Smith (2003) suggested that reliability might be more helpful for evaluating classroom assessment systems if thought of as sufficiency. Again, this is an analogous problem faced by those charged with evaluating the technical adequacy of alternate assessment systems and is similar to the technical challenges raised by the use of performance-based assessments (Linn & Burton, 1994). Students are not presented with a single multiple-choice test where a simple reliability coefficient can be computed quite easily. Most alternate assessment systems include relatively few open-ended tasks that are often tailored to the individual student. This type of system is not what traditional reliability methods were designed to measure. Some states report inter-rater reliability statistics as one indicator of reliability for alternate assessments. Although reporting the consistency of scoring processes is valuable, reporting inter-rater agreement statistics as if they are reliability coefficients is misleading. We need to conceptualize traditional reliability criteria so that they make sense given the unique features of alternate assessments.

Why should we do this? IDEA 97 required that states (and districts) develop alternate assessments to ensure all students with disabilities could show what they know in the “general curriculum,” in the context of standards-based reform. The Title I reauthorization in 1994 (IASA) had formalized the standards-based reform efforts of the previous decade by requiring that states define what knowledge and skills all children should know and be able to do, and to assess the performance of all children on that content. For the first time in many states, education stakeholders had to come to consensus on what the results of good teaching and learning should be for students, and to define publicly the parameters of the “general curriculum.” The shift had dramatic effect on how to measure student achievement. Large-scale assessment theory and practice developed as a means to sort and select examinees along a common “ability” continuum (Shepard, 2000). Students were believed to be distributed along a normal distribution and tests were designed to help fulfill this assumption (Shepard, 2000). The criterion-referenced testing movement and the current iteration of standards-based reform have changed the assumptions – now tests have to measure student achievement against *a priori* criteria and schools are being held accountable to ensure that all students reach these pre-established standards (NCLB, 2001).

As states began rethinking their approach in this new criterion-referenced environment, they were also grappling with another implication of both IASA 1994 and IDEA 1997: all students were to be assessed. Our understanding of large-scale assessment over the twentieth century had been built around the principle of standardized administration, standardized tasks, and standardized scoring. Measurement methodology was built to fit the standardized world, and many students with disabilities didn’t fit that world. The historical exclusion rates of students with disabilities from large-scale assessment are well documented (McGrew, Thurlow, & Spiegel, 1993; Shriner & Thurlow, 1993; Thurlow, Wiley, & Bielinski, 2003). The traditional emphasis on standardization as essential for ensuring the technical adequacy of large-scale assessments is part of the reason these students were excluded. Historic low expectations for achievement for students with disabilities also contributed to the exclusion of these students and to the acceptance of—even insistence on—their exclusion (McGrew & Evans,

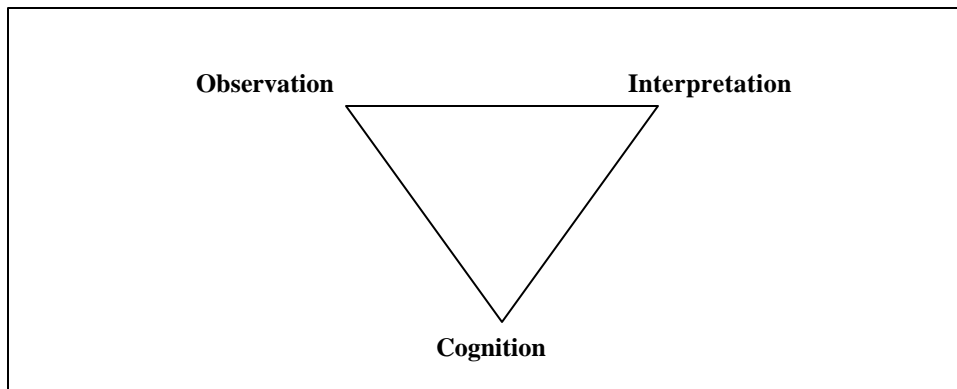
2004).

The flurry of concern about the effect of accommodations on assessment results evidenced in lawsuits in Indiana, Oregon, California, and now Alaska is one example of the aftermath of the inclusion of all children in standards-based instruction and in large-scale assessment of that instruction. This is playing out in legal arenas as well as educational ones (see, for example, Disability Rights Advocates, 2001). The push to develop understanding of “universally designed assessments” (Johnstone, 2003; Thompson, Johnstone, & Thurlow, 2002; Thompson, Thurlow, & Malouf, in press), which are cited in the NCLB Regulations, is also an example of the aftermath of the inclusion of students with disabilities in standards-based assessments.

There are pressures on traditional measurement models that have left measurement theorists and practitioners in a challenging situation, with measurement assumptions that don’t seem to ‘fit’ as well as they once did (Quenemoen & Marion, 2002). The option of removing students “who do not fit” from the population to resolve the dilemma is no longer an option. Assessments must fit all students, not the other way around.

There is a larger discussion occurring on whether current models of large-scale assessment appropriately reflect what we understand about what good teaching and learning looks like, and how students evidence that learning. (Pellegrino et al., 2001). Pellegrino et al. defined three pillars on which every assessment must rest: “a model of how students represent knowledge and develop competence in the subject domain, tasks or situations that allow one to observe students’ performance, and an interpretation method for drawing inferences from the performance evidence thus obtained.” (p. 2). They suggest that these three pillars make up an “assessment triangle,” and that this triangle—cognition, observation, and interpretation—must be articulated, aligned, and coherent in order for inferences drawn from the assessment to have integrity. They posit that it is the theory of learning—cognition—that is the “cornerstone” of the assessment design process. Figure 1 shows the triangle resting on the foundation of cognition, and building out to the observations and interpretation.

Figure 1: The assessment triangle (Pellegrino et al., 2001)



These authors suggest that as society is expecting more of traditional large-scale

assessments and requiring multiple uses of test results, we need to invest time and thought into improving how we “know what students know.” This can improve all forms of assessment in varying contexts and for varying purpose, whether classroom formative assessment or large-scale school accountability assessments. It provides an opportunity to ensure that assessment design processes build on understanding of how all students learn. It also requires attention to the need to understand if one learning theory fits all or whether some groups of students may represent knowledge and develop competence in the domain in somewhat or even dramatically different ways.

It is impossible to overestimate the challenge of rethinking a century of large-scale assessment tradition, along with the added complexity of rethinking how students learn and then show knowledge and skills in the content domains. By addressing the possibilities of new ways of thinking about knowing what students know for a group of children who have never been included in large-scale assessment for any purpose (e.g., students with the most significant cognitive disabilities), we believe we will be able to take a fresh look at where our traditions and conventions serve us well, and where they may not. By stepping away from what has become convention for general assessment, looking at the needs of a new population, we will discover hidden assumptions and issues in how we have been doing business all along, and define new directions to take us into the future.

Is this effort worth it? Given our apparent need to have a side-by-side glossary of assessment terms to translate our language in order to understand the issues, is it possible to work together to define new directions to take us into the future? And can we work together in ways that help us reconceptualize how all assessments can be improved?

A proposal for how we can work together to advance our mutual ability to build assessments that work for all students. The assessment triangle described by the NRC Committee on the Foundations of Assessment (Pellegrino et al., 2001) and discussed above can guide our work. Yet the Committee points out that “it is unlikely that the insights gained from current or new knowledge about cognition, learning, and measurement will be sufficient by themselves to bring about transformations in assessment... research and practice need to be connected more directly through the building of a cumulative knowledge base that serves both sets of interests” (p. 294).

To that end, Pellegrino et al. (2001) suggest that interdisciplinary partners from multiple communities should use the Committee’s conceptual scheme and language as a framework to guide improvement of current assessment materials, designs, and practices on the basis of existing knowledge. Simultaneously, these research and practice partnerships can yield new knowledge of how to conceptualize and operationalize assessments that result in more valid and fair inferences about student achievement in all areas of the school curriculum, for all children.

In working with cross-disciplinary research and practice partners thus far, we have identified essential research questions that include:

- Who are the learners who take alternate assessments? How does the type and size of the population vary in terms of learner characteristics, available response repertoires, and

complex medical conditions? How do the variations of who the learners are affect the assessment triangle, and ultimately technical adequacy studies?

- What does the literature say about how students in this population learn? How do current theories of learning for learners in the typical population apply to this population of students? How does this learning theory articulate with the assessment design, and ultimately with technical adequacy studies?
- How is technical adequacy defined? What is meant by reliability, validity? How do the traditional definitions of reliability/validity apply to alternate assessments? How do we define reliability and validity for different types of alternate assessments?
- What are the technical adequacy issues in alternate assessments that can not be resolved with the current knowledge-base in large-scale assessment? What strategies can be used to resolve the issues?
- What consequential validity issues (intended/unintended consequences) challenge the foundational assumptions in an alternate assessment? What is the relationship between foundational assumptions of alternate assessments and technical adequacy issues?
- What lessons learned from this study need to be addressed for the general assessment as well?

An essential first step in achieving this objective is to define the learners who take alternate assessments, and determine how these patterns differ across states. Students who typically participate in alternate assessments challenge the assessment triangle in that cognition in students from this population can only be observed through limited response repertoires. The type and size of the population is important from a technical adequacy point of view because within this one percent as defined in Title I Regulation (*Federal Register*, December 9, 2003) exists a highly variable population in terms of learner characteristics, available response repertoires, and often competing complex medical conditions.

Since the inception of alternate assessments a decade ago, the description of the population of students deemed eligible for alternate assessments ranged from students with severe and profound disabilities to some students with moderate disabilities. In most cases, these students represent less than 1% of the total population assessed in a large-scale assessment. For example, Kentucky (which has the longest history and most stable participation rate), assesses .8% of the total population; of those only about .4% of the scores would count as proficient. However, with the Title I one percent rule, the population may become broader and even more diverse. This is particularly true in states that have more than one alternate assessment; in 2003, eleven states indicated that they had multiple alternate assessments (Thompson & Thurlow, 2003). In these cases, it is likely that the type of learner will overlap in the various alternate assessments.

Second, it is essential to build consensus on a theory of learning in the academic content domains for alternate assessment participants. The literature on academic content learning for this population is limited and varied. As a field, we have not as yet grappled with a theory of learning in the academic content areas for these children, that is, what patterns of growth they show on the path to competence. Yet, these discussions have implications for content alignment and content extension discussions, discussions on assessment methods, scoring criteria, scoring processes, and standard-setting methods.

Finally, we need to step out of our specializations and think together about these challenges. In Appendix A, we provide a draft technical manual table of contents for alternate assessment of students with significant cognitive disabilities. Over the next five years, we hope to refine, change, or expand on our understanding of what would go into these chapters, develop understanding on how it differs from or improves upon current practice in documentation of large-scale general assessments, and ultimately build consensus on the criteria that can be used to judge technical quality of all assessments. We need all the partners at the table, learning each other's languages, and improving how we know what *all* students know.

References

- AERA/APA/NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education.
- Arnold, N. (2003). *Washington alternate assessment system technical report on standard setting for the 2002 portfolio* (Synthesis Report 52). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Browder, D. (2001). *Curriculum and assessment for students with moderate and severe disabilities*. New York: Guilford Press.
- Browder, D., Flowers, C., Ahlgrim-Delzell, L., Karvonen, M., Spooner, F., & Algozzine, R. (2002). *Curricular implications of alternate assessments*. Paper presented at the National Council of Measurement in Education Annual Conference, New Orleans.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.
- Disability Rights Advocates. (2001). *Do no harm – High stakes testing and students with learning disabilities*. Oakland, CA: Author.
- Federal Register*. (December 9, 2003).
- Freed, M. N., Hess, R. K., & Ryan, J. M. (2002). *The Educator's Desk Reference: A sourcebook of educational information and research*, 2nd edition. Oryx Press.
- Garrett B., Towles, E., Kleinert, H., & Kearns, J.F. (2003). Portfolios in large-scale alternate assessment systems: Frameworks for reliability. *Assessment for Effective Intervention*, 28 (2), 17-28.
- Johnstone, C.J. (2003). *Improving validity of large-scale tests: Universal design and student performance* (Technical Report 37). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Kearns, J. F., Kleinert, H., & Kennedy, S. (1999). Standards and assessments for all students - we need not exclude anyone! *Educational Leadership*, 56 (6), 33-38.
- Kleinert, H., & Kearns, J. (2001). *Alternate assessment: Measuring outcomes and supports for students with disabilities*. Baltimore: Brookes Publishing.
- Kleinert, H., & Kearns, J. (1999). A validation study of the performance indicators and learner outcomes of Kentucky's alternate assessment for students with significant disabilities. *Journal of The Association for Persons with Severe Handicaps*, 24(2), 100-110.

- Kleinert, H. L., Kearns, J.F., Kennedy, S. (1997). Accountability for all students: Kentucky's alternate portfolio assessment for students with moderate and severe disabilities. *The Journal of the Association for Persons with Severe Handicaps* 22(2), 88-101.
- Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex performance-based assessment: Expectations and validation criteria. *Educational Researcher*, 20, 8, 15-21.
- Linn, R. L. & Burton, E. (1994). Performance-Based Assessment: Implications of Task Specificity. *Educational Measurement: Issues and Practice*, 13, 5-8, 15.
- McGrew, K.S., & Evans, J. (2004). *Expectations for students with cognitive disabilities: Is the cup half-empty or half-full? Can the cup flow over?* (Powerpoint for draft paper). Available at www.iapsych.com/expect.files/frame.htm.
- McGrew, K.S., Thurlow, M.L., & Spiegel, A.N. (1993). An investigation of the exclusion of students with disabilities in national data collection programs. *Educational Evaluation and Policy Analysis*, 15 (3), 339-352
- Messick, S. (1995). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23, 2, 13-23.
- Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, 62, 229-258.
- Olson, B., Mead, R., & Payne, D. (2002). *A report of a standard setting method for alternate assessments for students with significant disabilities* (Synthesis Report 47). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.
- Quenemoen, R., Massanari, C., Thompson, S., & Thurlow, M. (2000). *Alternate assessment forum: Connecting into a whole*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Quenemoen, R. & Marion, S. (2003). *Rethinking basic assumptions of test development: Assessment frameworks for inclusive accountability tests* (Policy Directions No. 17). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Quenemoen, R., Rigney, S., & Thurlow, M. (2002). *Use of alternate assessment results in reporting and accountability systems: Conditions for use based on research and practice* (Synthesis Report 43). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Quenemoen, R., Thompson, S. & Thurlow, M. (2003). *Measuring academic achievement of students with significant cognitive disabilities: Building understanding of alternate assessment scoring*

- criteria* (Synthesis Report 50). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Roeber, E. (2002). Setting standards on alternate assessments (Synthesis Report 42). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Ryan, J.M., Quenemoen, R.F. & Thurlow, M.L. (2004). *I say potato and you say potahto: The assessment-speak gap between general and alternate assessment experts. A side-by-side glossary*. American Educational Research Association annual meeting presentation.
- Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of Research in Education, 19*, 405-450.
- Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, 29, 7, 4-14.
- Shriner, J.G., & Thurlow, M.L. (1993). *1992 State special education outcomes*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Thompson, S.J., Quenemoen, R., Thurlow, M.L., & Ysseldyke, J.E. (2001). *Alternate assessments for students with disabilities*. Thousand Oaks, CA: Corwin Press.
- Thompson, S.J., & Thurlow, M.L. (2001). *2001 State special education outcomes: A report on state activities at the beginning of a new decade*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Thompson, S., & Thurlow, M. (2003). *2003 State special education outcomes: Marching on*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Thompson, S., Thurlow, M., & Malouf, D. (in press). Creating better tests for everyone through universally designed assessments. *Journal of Applied Testing Technology*.
- Thurlow, M., Olsen, K., Elliott, J., Ysseldyke, J., Erickson, R., & Ahearn, E. (1996). *Alternate assessments for students with disabilities* (Policy Directions No. 5). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Thurlow, M.L., Wiley, H.I., & Bielinski, J. (2003). *Going public: What 2000-2001 reports tell us about the performance of students with disabilities* (Technical Report 35). University of Minnesota, National Center on Educational Outcomes.
- Tindal, G. (in press). *Alignment of Alternate Assessments Using the Webb System*. (Commissioned by CCSSO Technical Issues in Large Scale Assessment (TILSA) SCASS.

- Turner, M., Baldwin, L., Kleinert, H., & Kearns, J. (2000). An examination of the concurrent validity of Kentucky's alternate assessment system. *Journal of Special Education, 34*(2), 69-76.
- White, M., Garrett, B., Kearns, J., & Grisham-Brown, J. (2004). Instruction and Assessment: How students with deaf-blindness fare in large-scale alternate assessments. *Research and Practice for Persons with Severe Disabilities, 28* (4), 205-213.
- Wiener, D. (2002). *Massachusetts: One state's approach to setting performance levels on the alternate assessment* (Synthesis Report 48). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Ysseldyke, J. E., & Olsen, K. R. (1997). *Putting alternate assessments into practice: What to measure and possible sources of data* (Synthesis Report No. 28). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Appendix A. Draft Technical Manual Table of Contents

Recommendations for Alternate Assessments on Alternate Achievement Standards

The technical manual for any large-scale assessment provides information about the technical quality of assessments. The manuals typically include information on how the assessment was developed, administered, scored, and reported, as well as additional detail about any technical studies done on the completed assessment. The technical manual is an essential piece of evidence states can use to demonstrate the adequacy of their assessment system for Title I purposes.

A technical manual for alternate assessment on alternate achievement standards should have the following components.

Section I—Assessment Development

A. Overview

- Principles guiding development
- Partners and process guiding development
- Research base on desired outcomes for this population
- Documentation of process and result of state expansion/extension of the state content standards at grade level to ensure strong basis in literacy and numeracy
- Pros and cons of alternative methods considered
- Description of selected approach

B. Test Development

- Protocol for alignment to grade level content standards
- Development of draft assessment protocol
- Pilot test design and results
- Field test design and results

C. Test blueprint

- English Language Arts content specifications (see construct discussion above)
- Mathematics content specifications
- Other (e.g., Science) content specifications

Section II—Test Administration

A. Procedures for alternate assessment administration

- Decision-making process (participation, IEP team role)
- Local responsibility
- Timelines

B. Training

- Test oversight training for administrators
- Educator training for those working directly with students
- Ethical test administration training

Section III— Scoring and Reporting**A. Scoring design**

- Quality control
- Benchmarking
- Selecting and training scorers
- Scoring activities
- Inter-scorer reliability

B. Standard-setting

- Documented and validated process used for standard setting (Full description in Appendix _)
- Performance level descriptors and exemplars for alternate achievement standards
- Distribution of performance across levels
- Comparison of performance across levels achieved in general assessment

C. Reporting design

- School/District/State Report
- Parent Letter/Individual Student Report

Section IV - Reliability and Validity; Other Technical Considerations**A. Summary of studies for reliability, available data****B. Summary of studies for validity, available data**

- Face validity studies
- Concurrent validity studies
- Consequential validity studies

C. Other technical considerations**Section V—Appendices**

Appendix A Documentation of development principles, partners, process, research base

Appendix B Documentation of training provided, attendance, quality control

Appendix C Documentation of scoring protocols, process, quality control

Appendix D Formal evaluation data if available

Appendix E Standard setting report

Appendix F References

NASDSE VIDEO CONFERENCE MATERIALS – NCEO PRESENTATION

1. COMPREHENSIVE RESOURCES ON ACTUAL LAW, REGULATION, GUIDANCE DOCUMENTS FROM ed.gov

<http://www.ed.gov/policy/landing.jhtml?src=rt>

2. TECHNICAL TOOLS TO CONSIDER AS YOU THINK ABOUT TECHNICAL ADEQUACY

*** Peer Review Guidance – an essential resource. A short form checklist is attached as a separate Word file, but the entire document is the definitive resource.

<http://www.ed.gov/policy/elsec/guid/saaprguidance.doc>

A checklist developed by MSRRRC for the CCSSO presession is a “short form” tool for the Peer Review Guidance, and can be downloaded at:

<http://www.ihdi.uky.edu/msrrc/Word%20Docs/Final-%20Critical%20Element%20Checklist%20for%20Peer%20Review%20Guidance%20611.042.doc>

MSRRRC. (2004). *State standards and assessments: Critical elements checklist. Extracted from Standards and Assessments Peer Review Guidance: Information and Examples for Meeting the Requirements of NCLB (April 28, 2004).*

Quenemoen & Thurlow 2004 AERA Paper and Ryan, Quenemoen, & Thurlow 2004 AERA Glossary (attached in separate word files)

The paper includes a draft table of contents for an alternate assessment technical manual in the appendix and a reference list

NCEO Online Documents

NCEO Principles and characteristics of inclusive assessment and accountability systems:

<http://education.umn.edu/nceo/OnlinePubs/Synthesis40.html>

Thurlow, M., Quenemoen, R., Thompson, S., & Lehr, C. (2001). *Principles and characteristics of inclusive assessment and accountability systems* (Synthesis Report 40). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

And the accompanying self-study guide workbook:

<http://education.umn.edu/nceo/OnlinePubs/workbook.pdf>

Quenemoen, R. F., Thompson, S. J., Thurlow, M. L., & Lehr, C. A. (2001). *A self-study guide to implementation of inclusive assessment and accountability systems: A best practice approach*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.